Final Project — Open GWAS/Population Genetics Analysis

PUBH 8878

Purpose

- Analyze real genetic data using methods from the course. Choose one of the tracks below (or propose your own) and produce a succinct, reproducible analysis using a public dataset.
- This project emphasizes scientific reasoning, correct statistical practice, clear communication, and reproducible code. It is intentionally open-ended: depth > breadth.

Deliverables

- Project proposal (approx. 1 page): question(s), dataset(s), methods, risks/mitigations, expected outputs. **Due: Wednesday, October 15th by 11:59pm.**
- Final report (8–12 pages, PDF only): background and question, data, methods, results, limitations, references.
- Reproducible materials: a self-contained folder or repo with your `.qmd`/R scripts, environment notes, and seeds. Include a `README` with run instructions.
- Optional: 8–10 minute in-class or recorded talk with 2–3 slides of key results.

Evaluation (20% of course grade)

- Clarity and scope (20%): well-posed question; appropriate scope for time/resources.
- Methods and correctness (35%): sound statistical modeling, assumptions stated, correct inference, sensible QC.
- Reproducibility (20%): organized code; seeds; instructions; figures regenerate.
- Interpretation and communication (20%): figures/tables support claims; limitations; ethical awareness.
- Professionalism (5%): organization, writing, and citation quality.

Choose a Track (non-exhaustive)

- Track A — Secondary GWAS analysis (summary statistics): pick one trait and perform quality checks (QQ/λGC), Manhattan plot, locus zoom(s), gene/annotation enrichment, and short literature triangulation. Optional: SNP-heritability via LD Score Regression; cross-trait genetic correlation.
- Track B — Population structure and diversity: PCA/UMAP on a reference genotype panel; compute F_ST between populations; visualize allele frequency spectra; explore LD decay. Optional: ADMIXTURE/LEA ancestry components and interpretation caveats.
- Track C — Causal inference with two-sample Mendelian randomization: select a well-powered exposure/outcome with strong instruments; run multiple MR estimators; perform sensitivity and heterogeneity checks; discuss assumptions and violations.
- Track D — Fine-mapping or colocalization: focus on 1–2 loci; use LD from a reference panel; apply SuSiE/FINEMAP; or test GWAS–eQTL colocalization for a tissue of interest.
- Track E — Simulation with real LD: simulate phenotypes on a real genotype panel (e.g., chromosome 22) to study power, inflation, or PRS performance under different architectures.

Ethics & Responsible Use

- Use population labels with care; avoid essentialist interpretations. Discuss portability and fairness when comparing groups. Do not attempt re-identification. Respect each dataset's license/terms.

Data Sources (curated)

- GWAS Catalog (NHGRI–EBI): comprehensive registry of GWAS with summary statistics where available; good for trait curation and downloading per-study results. https://www.ebi.ac.uk/gwas/
- OpenGWAS (MRC IEU): programmatic access to >40k GWAS summary-stat datasets; integrates well with R packages `ieugwasr` and `TwoSampleMR`. https://gwas.mrcieu.ac.uk/
- Pan-UK Biobank (Broad): pan-ancestry GWAS results across thousands of phenotypes with interactive PheWeb and bulk download. https://pan.ukbb.broadinstitute.org/
- FinnGen: large disease-focused GWAS summary stats and phenotype documentation. https://www.finngen.fi/en/access_results
- Biobank Japan: GWAS results across many traits; multi-ancestry comparison opportunities. https://pheweb.jp/
- GIANT Consortium: anthropometric trait GWAS (e.g., height, BMI) — classic, clean testbeds. https://portals.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium
- Psychiatric Genomics Consortium (PGC): summary stats for psychiatric disorders; read and follow data use terms. https://www.med.unc.edu/pgc/download-results/

- 1000 Genomes Project (IGSR): open, phased whole-genome reference panel with population labels; ideal for PCA, F_ST, LD, and as an LD reference. https://www.internationalgenome.org/data
- HGDP + 1000G combined callset (gnomAD): harmonized WGS panel for global structure analyses (VCF/PLINK). https://gnomad.broadinstitute.org/downloads#v3-hgdp-1kg
- gnomAD v4: aggregated exome/genome allele frequencies; excellent for frequency-based analyses and QC (not individual-level genotypes). https://gnomad.broadinstitute.org/downloads
- GTEx/eQTL resources: eQTL Catalogue and GTEx v8 summary statistics for colocalization. https://www.ebi.ac.uk/eqtl/ and https://gtexportal.org/home/datasets
- LD reference (for LDSC/fine-mapping): precomputed 1000G LD scores and baseline annotations. https://data.broadinstitute.org/alkesgroup/LDSCORE/ and https://github.com/bulik/ldsc